



HAL
open science

Dataverse, an open and collaborative source research data repository software

Baptiste Rouxel, Alina Danciu, Jim Meyers, Geneviève Michaud, Tom Villette

► To cite this version:

Baptiste Rouxel, Alina Danciu, Jim Meyers, Geneviève Michaud, Tom Villette. Dataverse, an open and collaborative source research data repository software. 13th European DDI User Conference 2021, Training FAIR, EDDI, Nov 2021, Virtuel, France. pp.13. hal-03891601

HAL Id: hal-03891601

<https://hal-sciencespo.archives-ouvertes.fr/hal-03891601>

Submitted on 9 Dec 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution| 4.0 International License

Dataverse, an open and collaborative source research data repository software

Alina Danciu¹, Geneviève Michaud¹, Jim Myers², Baptiste Rouxel¹, Tom Villette¹

¹*Center for Socio-Political Data (CDSP)*

²*Global Dataverse Community Consortium (GDCC)*

—

“Dataverse is an open source software platform for sharing, finding, citing, and preserving research data (developed by the Data Science and Products team at the Institute for Quantitative Social Science and the Dataverse community).”

Source : <https://github.com/IQSS/dataverse>

What?

75 installations worldwide

Software installation, which then hosts multiple virtual archives called Dataverse collections. Each Dataverse collection contains datasets, and each dataset contains descriptive metadata and data files (including documentation and code that accompany the data). As an organizing method, Dataverse collections may also contain other Dataverse collections.

Why?

Dataverse is a good solution for networked institutions

- Single Sign On using institutional credentials
- Can harvest and be harvested
- General and Dataverse-level branding capabilities

Dataverse supports discoverability and persistence

- Powerful search engine and filters
- Good Google referencing
- Persistent identifiers

What for?

Nesstar => Dataverse

Harmonising metadata (institution names - in-house + ROR and ISNI, authors, keywords -ELSST...)

Implementing DDI controlled vocabularies

How?

In-house development of several scripts, some using Dataverse APIs:

- metadata harmonisation & conformity to CESSDA Metadata Model (CMM)
- migrating metadata from Nesstar to Dataverse

Installation and configuration of a Dataverse instance enabling:

- controlled vocabularies (according to the CMM, DDI CVs)
- multilingual UI (including CVs translation)
- OAI-PMH harvesting

Classification des sujets ⓘ

Terme ⓘ

Vocabulaire ⓘ

The screenshot shows a search interface with a dropdown menu for 'Terme' and an input field for 'Vocabulaire'. The dropdown menu is open, displaying a list of terms. The term 'Santé au travail' is highlighted in blue. The other terms in the list are: 'Services et politiques de soins', 'Médicaments et traitements', 'Condition et activité physique', 'Santé publique', 'Santé reproductive', and 'Signes et symptômes ; conditions pathologiques'.

Terme	Vocabulaire
Santé au travail	
Services et politiques de soins	
Médicaments et traitements	
Santé au travail	
Condition et activité physique	
Santé publique	
Santé reproductive	
Signes et symptômes ; conditions pathologiques	

data.sciencespo repository

CDSP data bank & institutional self deposit repository released January 2020:

- 7 custom CVs
- 300+ metadata items harmonised & migrated
- EN | FR user interface
 - including CVs translations
- custom homepage and style
- harvested by:
 - [CESSDA DATA Catalogue \(select French UI\)](#)
 - french SSH portal [Isidore](#)
 - European [OpenAIRE](#) portal



Dataverse communities

- Other groups working on similar issues ([PyDataverse](#), [SuperDADA](#) etc)
- CDSP's contributions to IQSS DV codebase (pull requests)
- Opening of several issues aimed at OAI-PMH endpoints
 - harvesting by the CESSDA Data Catalogue
- Decision to sponsor a list of key issues (GDCC contract and collaboration)
 - tackling **concerns shared among users in the DDI community**

<https://github.com/orgs/IQSS/projects/12>

The screenshot shows a GitHub Project board for the 'data.sciencespo - Sciences Po institutional research data repository'. The board is organized into three columns: 'Interest' (5 items), 'Needed' (1 item), and 'Done' (10 items). Each item is a card representing a task or issue, with details such as the issue number, title, assignee, and status.

Interest (5 items):

- Allow customization of Shibboleth redirects per institution** (dataverse#3911) opened by donsize more. Feature: Account & User Info. User Role: Depositor.
- Bootstrap 4 Upgrade** (dataverse#6045) opened by djbrooke. Large.
- As a Dataverse admin, I would like to hook up my installation to an IDMI/IAM** (dataverse#5974) opened by polkiotherm. 4 of 5 tasks.
- Dataverse doesn't always produce valid DDI codebook 2.5 XML** (dataverse#3648) opened by jomtov. Features: Harvesting, Metadata. Type: Bug. User Role: Sysadmin.
- Dataset Citation: provide flexible options for information displayed in citation metadata** (dataverse#2297) opened by sbarbosadataverse. Features: Metadata, Suggestion. User Role: Curator.

Needed (1 item):

- Shibboleth users who predate Shibboleth are assumed to be email-verified but aren't** (dataverse#5963) opened by donsize more. Small.

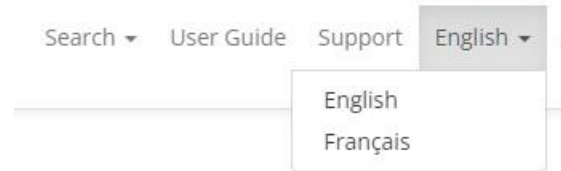
Done (10 items):

- Performance: Very sluggish page load when languages and all facets enabled in root dv.** (dataverse#6052) opened by kcondon. Features: Performance & Stability, Medium. 1 linked pull request.
- Docker for production** (dataverse#4666) opened by omarsaloudani. Features: Installer. Status: Still Interested? 1 linked pull request.
- The metadata tab of a dataset does not display the translated values** (dataverse#6607) opened by bappun. Feature: Internationalization. 1 linked pull request.
- As an admin and developer, I'd benefit from a simplified configuration mechanism** (dataverse#5293) opened by polkiotherm.
- Delete user via API** (dataverse#4475) opened by laertecchla. Features: Account & User Info, Large. User Role: Superuser. 5.4.

Internationalization in Dataverse

Viewing in multiple languages

- User Interface translations
- Translations of metadata fields with controlled vocabularies



Issues raised by Sciences Po re: Content Creation/Machine Harvesting:

- Metadata field to specify the language(s) used in data files, but
- No way to indicate the language used to enter metadata for a dataset
- Metadata exports (including the DDI export used for harvesting by CESSDA) do not indicate the language used

Specifying the Metadata Language

Admins enable/specify the allowed languages

Sub-collections are configured to indicate the metadata language allowed for new datasets

Dataset creators/editors are prompted to use the configured language

Dataset Metadata Language ?

fr (inherited from enclosing Data) ▼

English

Français

fr (inherited from enclosing Dataverse)

Metadata Fields

Please enter metadata in the **Français** language, or change the metadata language choice in the parent dataverse before proceeding.

Title ?

Test d'internationalisation

Alternative Title ?

Files Metadata Terms Versions

This dataset has been configured to use **English** as the language for all metadata entries.

Data Project Persistent ID ? doi:10.33564/FK2RS9JLQ

Publication Year ? 2021-09-16

Title ? 5.6 test

Exporting Language Information

Metadata exports include metadata language, i.e. DDI:

Overall xml:lang attribute

```
<codeBook xmlns="ddi:codebook:2_5"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="ddi:codebook:2_5
https://ddialliance.org/Specification/DDI-Codebook/2.5/XMLSchema/codebook.xsd"
version="2.5" xml:lang="fr">
```

Individual xml:lang attributes on specific fields

```
<abstract xml:lang="fr">TEST</abstract>
```

Controlled values exported in site default language and dataset's metadata language

```
<sumDscr>
  <anyUnit xml:lang="en">Family: Household family</anyUnit>
  <anyUnit xml:lang="en">Household</anyUnit>
  <anyUnit xml:lang="fr">Famille : ménage</anyUnit>
  <anyUnit xml:lang="fr">Ménage</anyUnit>
</sumDscr>
```

* Thanks for the [online CESSDA validator!](#) - used to identify where language information was needed/recommended

Community Processes



Open Source on GitHub

Community-contributed issues and code contributions

Community contributions to language translations



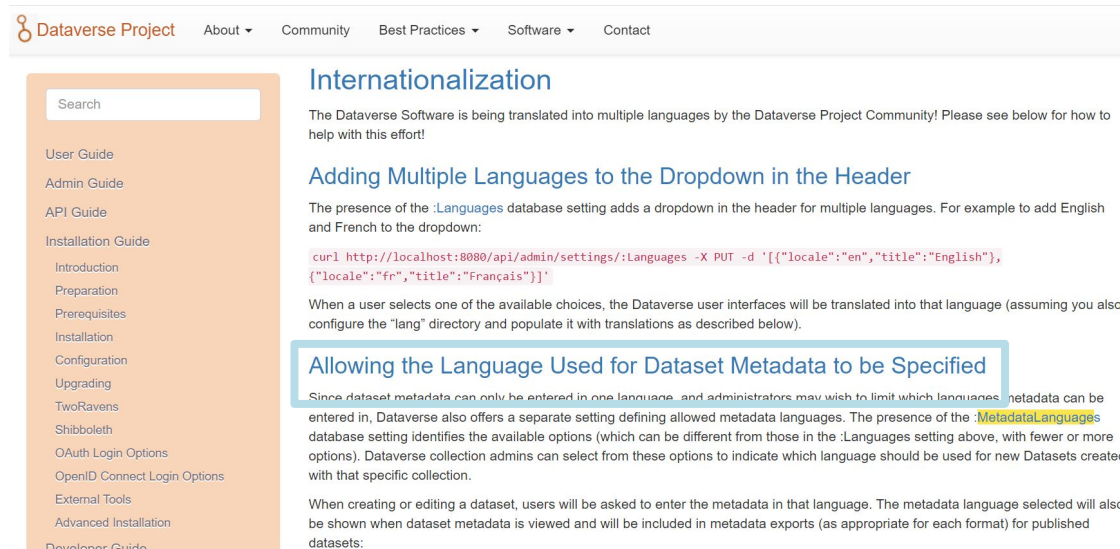
GDCC - Global Dataverse Community Consortium

Prioritizing work of interest across the community

Contracting to enable members to engage experienced developers/admins on their priorities

Dataverse 5.7+

New functionality for describing datasets/supporting harvesting by CESSDA in multiple languages thanks to Sciences Po!



The screenshot shows the Dataverse Project website. The navigation bar includes links for About, Community, Best Practices, Software, and Contact. A search bar is located in the top left. A sidebar on the left contains a search box and a list of navigation links: User Guide, Admin Guide, API Guide, Installation Guide, Introduction, Preparation, Prerequisites, Installation, Configuration, Upgrading, TwoRavens, Shibboleth, OAuth Login Options, OpenID Connect Login Options, External Tools, Advanced Installation, and Developer Guide. The main content area features a section titled "Internationalization" with a sub-section "Adding Multiple Languages to the Dropdown in the Header". This sub-section includes a code block for a curl command and a text block explaining the configuration. A blue-bordered box highlights the sub-section "Allowing the Language Used for Dataset Metadata to be Specified", which contains text about limiting metadata languages. The footer of the page is partially visible at the bottom.

[Dataverse Project](#) [About](#) [Community](#) [Best Practices](#) [Software](#) [Contact](#)

Search

- User Guide
- Admin Guide
- API Guide
- Installation Guide
 - Introduction
 - Preparation
 - Prerequisites
 - Installation
 - Configuration
 - Upgrading
 - TwoRavens
 - Shibboleth
 - OAuth Login Options
 - OpenID Connect Login Options
 - External Tools
 - Advanced Installation
 - Developer Guide

Internationalization

The Dataverse Software is being translated into multiple languages by the Dataverse Project Community! Please see below for how to help with this effort!

Adding Multiple Languages to the Dropdown in the Header

The presence of the `.Languages` database setting adds a dropdown in the header for multiple languages. For example to add English and French to the dropdown:

```
curl http://localhost:8080/api/admin/settings/.Languages -X PUT -d '{"locale":"en","title":"English"}, {"locale":"fr","title":"Français"}'
```

When a user selects one of the available choices, the Dataverse user interfaces will be translated into that language (assuming you also configure the `"lang"` directory and populate it with translations as described below).

Allowing the Language Used for Dataset Metadata to be Specified

Since dataset metadata can only be entered in one language, and administrators may wish to limit which languages metadata can be entered in, Dataverse also offers a separate setting defining allowed metadata languages. The presence of the `.MetadataLanguages` database setting identifies the available options (which can be different from those in the `.Languages` setting above, with fewer or more options). Dataverse collection admins can select from these options to indicate which language should be used for new Datasets created with that specific collection.

When creating or editing a dataset, users will be asked to enter the metadata in that language. The metadata language selected will also be shown when dataset metadata is viewed and will be included in metadata exports (as appropriate for each format) for published datasets: